Quarterly Reviews of Biophysics 36, 3 (2003), pp. 307–340. © 2003 Cambridge University Press
DOI: 10.1017/S0033583503003901 Printed in the United Kingdom

Prediction of protein function from protein sequence and structure

James C. Whisstock[1] and Arthur M. Lesk[1,2]*

[1] Department of Biochemistry and Molecular Biology, Victorian Bioinformatics Consortium, Monash University, Clayton Campus, Melbourne, Victoria 3168, Australia and ARC Centre for Structural and Functional Microbial Genetics, Monash University, Clayton, Victoria, Australia
[2] Cambridge Institute for Medical Research, University of Cambridge Clinical School, Wellcome Trust/MRC Building, Hills Road, Cambridge, CB2 2XY, UK

Abstract. The sequence of a genome contains the plans of the possible life of an organism, but implementation of genetic information depends on the functions of the proteins and nucleic acids that it encodes. Many individual proteins of known sequence and structure present challenges to the understanding of their function. In particular, a number of genes responsible for diseases have been identified but their specific functions are unknown. Whole-genome sequencing projects are a major source of proteins of unknown function. Annotation of a genome involves assignment of functions to gene products, in most cases on the basis of amino-acid sequence alone. 3D structure can aid the assignment of function, motivating the challenge of structural genomics projects to make structural information available for novel uncharacterized proteins. Structure-based identification of homologues often succeeds where sequence-alone-based methods fail, because in many cases evolution retains the folding pattern long after sequence similarity becomes undetectable. Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. Many methods of function prediction rely on identifying similarity in sequence and/or structure between a protein of unknown function and one or more well-understood proteins. Alternative methods include inferring conservation patterns in members of a functionally uncharacterized family for which many sequences and structures are known. However, these inferences are tenuous. Such methods provide reasonable guesses at function, but are far from foolproof. It is therefore fortunate that the development of whole-organism approaches and comparative genomics permits other approaches to function prediction when the data are available. These include the use of protein–protein interaction patterns, and correlations between occurrences of related proteins in different organisms, as indicators of functional properties. Even if it is possible to ascribe a particular function to a gene product, the protein may have multiple functions. A fundamental problem is that function is in many cases an ill-defined concept. In this article we review the state of the art in function prediction and describe some of the underlying difficulties and successes.

* Author to whom correspondence should be addressed. A. M. Lesk, Cambridge Institute for Medical Research, University of Cambridge Clinical School, Wellcome Trust/MRC Building, Hills Road, Cambridge, CB2 2XY, UK. (E-mail: aml2@mrc-lmb.cam.ac.uk)

1. Introduction

Much of the evolution of living systems on the molecular level proceeds according to the cascade:

gene sequence determines amino-acid sequence,
amino-acid sequence determines protein structure,
protein structure determines protein function,
selection acts on function to modify allele frequencies in populations (to close the loop).

Genome sequencing projects produce the full DNA sequences of organisms. Identification of genes within genomes provides the amino-acid sequences of the organism's proteins. In structural genomics projects, X-ray crystallography and NMR spectroscopy aim to determine the structures of a subset of the proteins from which other structures can be predicted by homology modelling. Contemporary bioinformatics collects data on sequences, structures, and functions, and studies the correspondences between them (for general references, see Galperin & Koonin, 2002; Lesk, 2001, 2002).

Assignments of function are based either solely on amino-acid sequences (the most common situation that arises frequently in annotating newly sequenced genomes), on some combination of sequence and structure, or on some organism-wide data, if available, such as protein–protein

interaction patterns (von Mering *et al.* 2003). The problem of predicting protein function arises in two general contexts. (1) The interest of a research group may be focussed on a gene and its protein product, and the group may pursue its investigation in detail; such studies may include identification of cofactors and post-translational modifications, and even a structure determination and a check of the phenotypic effect of a knockout. The result is an attempt to assign function on the basis of a thick dossier of detailed information. In the past this was the paradigm. (2) However, with increasing frequency we must deal with much sparser information. The largest sources of proteins of unknown function are complete genome sequences, giving us the challenge of annotating them (Smith, 1998; Eisenberg *et al.* 2000; Stein, 2001; Thornton, 2001). In these cases the data about specific proteins in genomes are often limited to their amino-acid sequences. The goal of providing at least approximate structural information, for its implications about function, is an important motivation of structural genomics projects (Burley *et al.* 1999; Eisenstein *et al.* 2000; Skolnick *et al.* 2000; Brenner, 2001; Chance *et al.* 2002; Gilliland *et al.* 2002; Zhang & Kim, 2003).

In analysing a novel genome, how well do we understand Nature's rules in proceeding from DNA sequence to amino-acid sequence to protein structure to function?

- Starting from a genome sequence, gene identification is still problematic, especially in eukaryotes where alternative splicing patterns compound the difficulty (Novichkov *et al.* 2001; Jones *et al.* 2002).
- The next step is perhaps the safest: Based originally on the experiments of Anfinsen demonstrating the reversible denaturation of proteins, we know that Nature has strict rules for determining protein structure uniquely from amino-acid sequences. There are a few exceptions – notably the prion proteins (Cohen & Prusiner, 1998; Peretz *et al.* 2002), and the serpins (Whisstock *et al.* 1998; Gettins, 2002; Pike *et al.* 2002) but this generalization is among the most robust we have in the field. (Chaperones are only catalytic in this process, not containing any information specific to the folding of any particular protein.) Although as yet we do not understand the physical basis of Nature's folding algorithm in sufficient detail to predict structure from sequence, progress is being made (Schonbrun *et al.* 2002; Tramontano, 2003). Moreover, the observation that similar sequences determine similar structures (the 'differential form' of the folding problem) gives us general confidence in homology modelling. Much less reliable is the widely held assumption that proteins with very similar sequences should – by virtue of their very similar structures – have similar functions.
- To reason from sequence and structure to function is to step onto much shakier ground. Following the reasoning of the previous paragraph, a common way to try to assign function to a protein is to identify a putative homologue of known function and guess that both share a common function. It is indeed true that many families of proteins contain homologues with the same function, widely distributed among species; for these, reasoning from homology does assign function correctly. However, the assumption that homologues share function is less and less safe as the sequences progressively diverge. Moreover, even closely related proteins can change function, either through divergence to a related function or by recruitment for a very different function (Ganfornina & Sánchez, 1999). In such cases, assignment of function on the basis of homology, in the absence of direct experimental evidence, will give the wrong answer, leading to misannotations in databanks. Many authors have called attention to 'howlers' in annotation (Smith & Zhang, 1997; Bork *et al.* 1998; Bork & Koonin, 1998; Doerks *et al.* 1998; Karp, 1998; Brenner, 1999; CODATA Task Group, 2000;

Devos & Valencia, 2000, 2001; Gerlt & Babbitt, 2000; Jeong & Chen, 2001). Iyer *et al.* (2001) have collected cases in which prediction and experiment agree, but both are likely to be wrong!

Indeed, the situation can be even worse. An often-asked question is: 'How much must a protein change its sequence before its function changes?' The answer is: 'Not at all!' There are numerous examples of proteins with multiple functions:

(1) Eye lens proteins in the duck are identical in sequence to active lactate dehydrogenase and enolase in other tissues, although they do not encounter the substrates in the eye. They have been recruited to provide a completely unrelated function based on the optical properties of their assembly. Several other avian eye lens proteins are identical or similar to enzymes. In some cases residues essential for catalysis have mutated, proving that the function of these proteins in the eye is not an enzymic one (Wistow & Piatigorsky, 1987). Note that the coexistence in some species of mutated inactive enzymes in the eye, and active enzymes in other tissues, implies that the gene must have been duplicated.

(2) Certain proteins interact with different partners to produce oligomers with different functions. In *Escherichia coli*, a protein that functions on its own as lipoate dehydrogenase is also an essential subunit of pyruvate dehydrogenase, 2-oxoglutarate dehydrogenase and the glycine cleavage complex (Riley, 1997).

(3) Proteinase do functions as a chaperone at low temperatures and as a proteinase at high temperatures. The logic, apparently, is that under conditions of moderate stress it attempts to salvage misfolded proteins; under conditions of higher stress it 'gives up' and recycles them (Spiess *et al.* 1999).

(4) Phosphoglucose isomerase (=neuroleukin=autocrine motility factor=differentiation and maturation mediator) functions as a glycolytic enzyme in the cytoplasm, but as a nerve growth factor and cytokine outside the cell (Jeffery, 1999; Jeffery *et al.* 2000). The structural origin of the extracellular receptor function is obscure.

These cases imply that *even if* detailed studies of the classical biochemical type on isolated proteins in dilute salt solutions do identify a function, we cannot be sure that we know the molecule's full repertoire of biological activities.

Conversely, non-homologous proteins may have similar functions. Chymotrypsin and subtilisin, two proteinases that even share a common Ser-His-Asp catalytic triad, are not homologous, and show entirely different folding patterns (Fig. 1). They are a standard example of convergent evolution. The Ser-His-Asp triad also appears in other proteins, including lipases and a natural catalytic antibody. This and other examples show that it is not possible to reason that if two proteins have different folding patterns they must have different functions.

In summary (see Fig. 2):

- *Similar sequences produce similar protein structures*, with divergence in structure increasing progressively with the divergence in sequence (Chothia & Lesk, 1986).
- Conversely, *similar structures are often found with very different sequences.* For instance, many proteins form TIM barrels with no easily detectable relationship between their sequences (Copley & Bork, 2000; Nagano *et al.* 2002).
- *Similar sequences and structures sometimes produce proteins with similar functions*, but exceptions abound (Ponting, 2001; an extensive table appears in Rost, 2002).
- Conversely, *similar functions are often carried out by proteins with dissimilar structures*; examples include the many different families of proteinases, sugar kinases, and lysyl-tRNA synthetases (Doolittle, 1994; Galperin *et al.* 1998).

Fig. 1. Ch
patterns, th
function an

Fig. 2. Or
similar seq
quences. Bi
solid outlin

Because ev
function re
specify fur
barrel is li
investigati
structure,

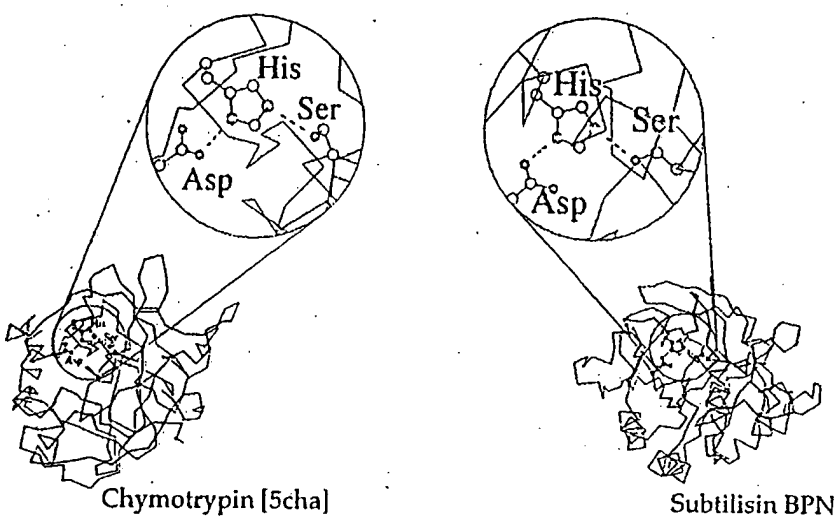


Fig. 1. Chymotrypsin and subtilisin are both proteinases. Although they have entirely different folding patterns, they share a common mechanism, including the catalytic triad Ser-His-Asp. The similarity of function and mechanism has arisen by convergent evolution.

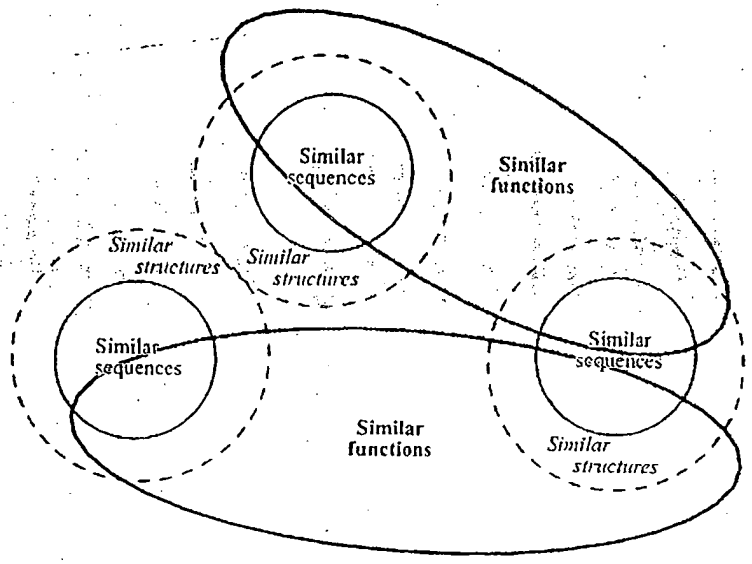


Fig. 2. Organization of the spaces of protein sequences, structures and functions. Thin solid outlines: similar sequences produce similar structures, but not all similar structures have recognizably similar sequences. Broken outlines: proteins of different sequence and structure can share similar functions. Thick solid outlines: conversely, proteins of similar sequence and structure can show different functions.

Because evolution has so assiduously pushed the limits in its exploration of sequence–structure–function relationships, many procedures described in the literature on function prediction do not specify function exactly, but do provide general hints. For instance, a protein known to be TIM barrel is likely to be a hydrolytic enzyme. Such hints are very useful in guiding experimental investigations of function, and indeed a sufficient accumulation of hints – based on sequence, structure, genomics, and interaction patterns – may well allow an expert to make a reasonable

proposal of a specific function. However, such an approach, relying as it does on human expertise, is difficult to automate for high-throughput full-genome analysis.

Two examples from the *Haemophilus influenzae* structural genomics project illustrate the point. High-resolution crystal structures of the proteins HI1434 (Zhang *et al.* 2000) and of HI1679 (Parsons *et al.* 2002) have been determined. Hypothesis of function from evolutionary relationships and detailed examination of the structures, followed by experimental verification of correct functional assignment, was successful in the case of HI1679 but until now have not yet proved successful for HI1434.

HI1679 has an α/β-hydrolase fold, with putative remote homology, based on sequence analysis, to members of the L-2-haloacid dehydrogenase family, the P-domain of Ca^{2+} ASPase and phosphoserine phosphatase. It was the first structure of a protein in the L-2-haloacid dehydrogenase family to be determined, and one of the motives for selecting it for investigation was the goal of learning about the structure and the mechanism of function of this family. The structure was consistent with a phosphatase, and this was confirmed by trying a variety of potential substrates. The protein cleaved 6-phosphogluconate and phosphotyrosine, confirming it to be a phosphatase. Addressing the original goal of elucidating the functions of this family of proteins, observed substrates were modelled into the binding pocket to supply suggestions about how sequence variation in the active site might affect specificity (Parsons *et al.* 2002).

HI1434 is related to a region in tRNA synthetases. The structure showed a putative binding site, a cleft that was conserved in the modelled structures of homologues. The structure itself and its evolutionary relationships suggest that it binds a nucleotide in its cleft. However, in this case no specific ligand has so far been identified.

View it in these terms: Inferring protein function from knowledge of the function of a close homologue is like solving the clue of an American crossword puzzle. Finding the word that satisfies the definition may be difficult but the task is in principle straightforward. Working out the function of a protein from its sequence and structure is like solving the clue of a British crossword puzzle. It is by no means obvious which features of the definition are providing the real clues, as opposed to misleading ones. Also, for both types of puzzle and for the suggestion of a protein function, even if your answer appears to fit it may be wrong.

2. Plan of this article

Our goal is to review methods that have been proposed for prediction of protein function from amino-acid sequence and three-dimensional (3D) structure, and, as far as possible, to evaluate them. However, it is difficult to state criteria for successful prediction of function, since function is in principle a fuzzy concept. Given three *sequences*, it is possible to decide which of the three possible pairs is the most closely related. Given three *structures*, methods are also available to measure and compare the similarity of the pairs. However, in many cases, given three protein *functions*, it would be more difficult to choose the pair with the most similar function. For example, although it is possible to define metrics for quantitative comparisons of different protein sequences and structures, this is more difficult for different protein functions.

Comparisons of functions could be based on suggested classifications of functions. There are many such classifications (recently reviewed by Ouzounis *et al.* 2003). Probably the most widely known is the Enzyme Commission (EC) scheme, limited of course to that class of functions. Other protein function classification schemes have been proposed, many in connection with individual organisms or individual families of proteins. However, a scheme appropriate for one

organism is not necessarily appropriate for others, and until recently there has been no noticeable attempt at consistency.

Indeed, even for very well understood proteins, there are different legitimate points of view about what aspects of function to focus on. The biochemist looks for the process mediated by the isolated protein in dilute solution. The molecular biologist looks for the significance, in the overall scheme of the life of the cell, of the process or processes in which the protein participates. We describe and compare various schemes for classifying protein function and ask whether it is possible to reconcile the different points of view. We also suggest that the Gene Ontology Consortium offers the most attractive approach.

If we had a classification of protein functions, we would want to map it onto classifications of sequence and structure. Classifications of sequence and structure are available, based on evolutionary principles. Therefore, to work with an appropriate classification of function, it is useful to understand how evolving proteins develop different functions. After developing this as background, we describe and classify the various methods that have been used to predict protein function and annotate genomes.

3. Natural mechanisms of development of novel protein functions

Information available about how proteins alter existing functions or develop new ones is abundant, although most of it is more anecdotal than systematic. Observed mechanisms of protein evolution that produce altered or novel functions include: (1) Divergence, (2) Recruitment and (3) 'Mixing and matching' of domains.

3.1 Divergence

In families of closely related proteins, mutations usually conserve function but modulate specificity. For example, the trypsin family of serine proteinases contains a specificity pocket: a surface cleft complementary in shape and charge distribution to the side-chain adjacent to the scissile bond. Mutations tend to leave the backbone conformation of the pocket unchanged but to affect the shape and charge of its lining, altering the specificity.

The change in specificity of the proteases illustrates a common theme: Although homologous proteins show a general drifting apart of their sequences as they accumulate mutations, often a few specific mutations account for functional divergence (Golding & Dean, 1998), as initially proposed by Perutz (1983) for haemoglobin. The malate and lactate dehydrogenase (MDH/LDH) family is a good example. Malate and lactate dehydrogenases are related enzymes catalysing related reactions. Wilks *et al.* (1988) showed by site-directed mutagenesis that a single residue change could switch the activity. Their paper may have been read by a trichomonad, which developed an MDH that, in a family tree of these enzymes, is much more similar to LDH molecules than to other MDHs, and appears to have arisen by convergent evolution (Wu *et al.* 1999).

The TIM-barrel structure, or very similar variants, has now appeared in over 100 enzymes of known crystal structure (Fig. 3). In many cases the sequence similarity is so low that it is impossible to say whether the proteins are genuinely related, or whether evolution has discovered this very stable and useful fold more than once. Conversely, certain enzymes sharing the TIM-barrel fold, and which are similar enough for us to be confident of their homology, clearly show the divergent evolution of new functions (Copley & Bork, 2000; Anantharaman *et al.* 2003).

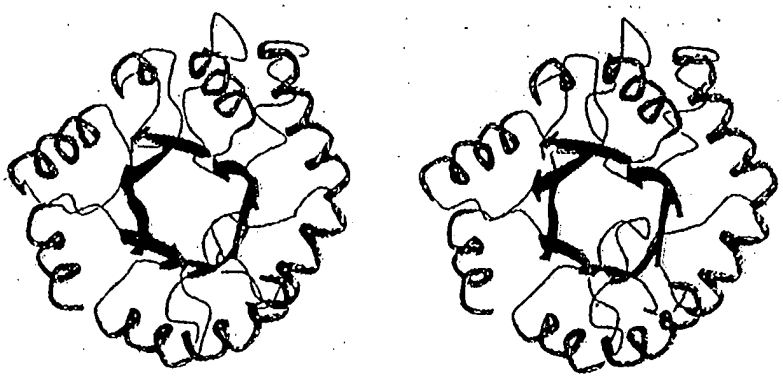


Fig. 3. Spinach glycolate oxidase, one of the many enzymes with the TIM-barrel structure. In this view into the barrel, the molecule is orientated with the C-termini of the strands nearer the viewpoint. This is the side of the barrel that in most cases carries the active site.

The enolase superfamily, which exhibits a folding pattern very closely related to TIM-barrel, contains several enzymes that catalyse different reactions with shared features of their mechanisms (Hasson *et al.* 1998). These include enolase itself, mandelate racemase, muconate lactonizing enzyme I, and D-glucarate dehydratase. From the point of view of sequence similarity, these enzymes are fairly close relatives. Mandelate racemase and muconate lactonizing enzyme I have 25% sequence identity. However, looking only at sequence and structure runs the risk of overlooking a more subtle similarity. What these enzymes share is a common feature of their *mechanism*. Each acts by abstracting a proton adjacent to a carboxylic acid to form an enolate intermediate (Fig. 4). The stabilization of a negatively charged transition state is conserved. In contrast, the subsequent reaction pathway, and the nature of the product, vary from enzyme to enzyme. These enzymes have not only a similar overall structure, a variant of the TIM-barrel fold, but each requires a divalent metal ion, bound by structurally equivalent ligands. Different residues in the active site produce enzymes that catalyse different reactions.

An aspect of divergence important for its implications about function is the distinction between orthologues and paralogues. Any two proteins that are related by descent from a common ancestor are homologues. Two proteins in different species descended from the same protein in an ancestral species are orthologues. Two proteins related via a gene duplication within one species (and the respective descendants of the duplicates) are paralogues. After gene duplication, one of the resulting pairs of proteins can continue to provide its customary function, releasing the other to diverge to develop new functions. Therefore inferences of function from homology are more secure for orthologues than for paralogues.

The database, Clusters of Orthologous Groups (COGs), is a collection of proteins encoded in fully sequenced genomes, organized into families (Natale *et al.* 2000). The COGs database has been applied to analysis of function and genome annotation.

Comparative analyses of known structures in such families of enzymes illustrate the kinds of structural features that change and those that stay the same. In some cases, the catalytic atoms occupy the same positions in molecular space, although the residues that present them are located at different positions in the sequence. In other cases the positions in space of the catalytic residues are conserved even though the identities and functions of the catalytic residues vary. In these cases, there appears to be a set of conserved 'functional positions' relative

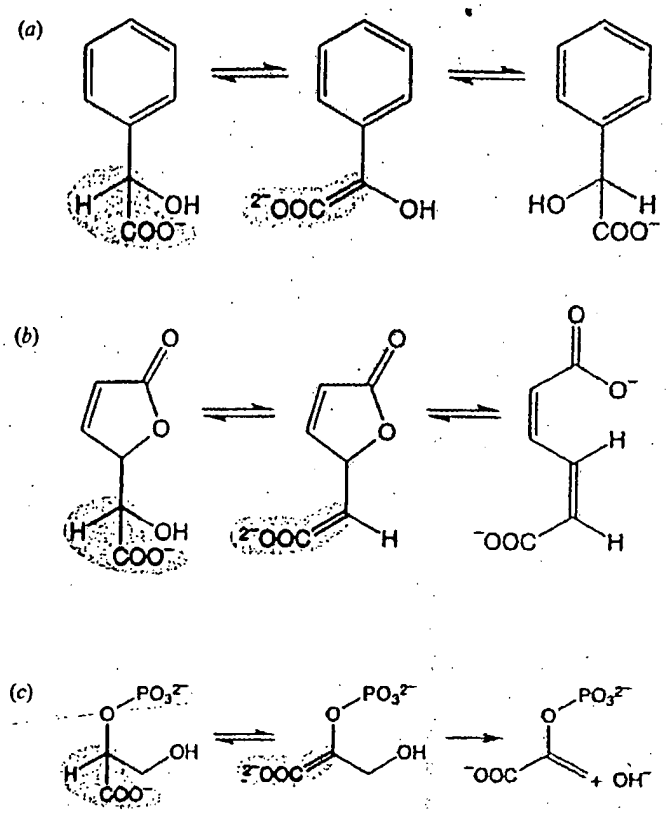


Fig. 4. Common mechanism in the enolase family of enzymes: (*a*) mandelate racemase, (*b*) muconate lactonizing enzyme, (*c*) enolase. (After Hasson *et al.* 1998.)

to the molecular framework. When functional residues are conserved in this way, in the structure if not necessarily in the sequence, they can provide a signal from which we can recognize function.

However, several enzyme families show an even greater degree of divergence, including variation in the residues responsible for mediating catalysis. For example, the Apurinic/Apyrimidinic endonuclease superfamily is a large diverse family of phosphoesterases. The family includes members that cleave nucleic acids (both DNA and RNA). However, the family has diverged to include lipid phosphatases. The essential catalytic residues vary between different subfamilies, for example, an essential His in the DNA repair enzyme DNaseI is not conserved in exonuclease III. In these cases, the conservation patterns from which we could hope to identify function have disappeared.

In some cases very large divergence has led to very different function. Murzin (1998) and Grishin (2001) have discussed how far divergence can push the relationships between homology, structure and sequence divergence, and functional change. Some changes in folding pattern, or topology, associated with functional changes, are:

(1) Addition/deletion/substitution of secondary structural elements. A dramatic example is the relationship between luciferase and a non-fluorescent flavoprotein, which, although they have 30% sequence identity show a standard TIM barrel in the case of luciferase but a truncated barrel in the non-fluorescent flavoprotein.

(2) Circular permutation. An example is NK-lysin, an all-α protein, and an aspartic proteinase prophytepsin.
(3) Stand invasion and withdrawal. Although insertion of strands at the end of a β-sheet is relatively simple, it is more difficult to insert a strand into a β-barrel. Lipocalins include the homologues retinol-binding protein with an 8-stranded β-barrel, and retinoic acid-binding protein with a 10-stranded β-barrel.
(4) Changing the topology while maintaining the architecture. *Aeromonas* aminopeptidase and carboxypeptidase G_2 have a common core of secondary structural elements, but are 'wired up' by connecting loops in a different way. The thrombin inhibitor triabin is likely to be related to the lipocalins, on the basis of similarities in the amino-acid sequences. Both contain β-barrel folds, but superposing the structures shows that two of the strands have been swapped.

3.2 Recruitment

The application of enzymes as lens crystallins illustrated another route of evolution: a novel function *preceding* divergence. It is more difficult to distinguish divergence and recruitment than it might first appear. Divergence and recruitment are at the ends of a broad spectrum of changes in sequence and function. Apart from cases of 'pure' recruitment such as the duck eye lens proteins or phosphoglucose isomerase, in which a protein adopts a new function with no sequence change at all, there are examples, not only of relatively small sequence changes correlated with very small function changes (which most people would think of as relatively pure divergence), and relatively small sequence changes with quite large changes in function (which most people would think of as recruitment), but also many cases in which there are large changes in both sequence and function.

3.3 'Mixing and matching' of domains, including duplication/oligomerization, and domain swapping or fusion

Many large proteins contain tandem assemblies of domains which appear in different contexts and orders in different proteins. (The reader must be aware that there is no universal agreement about how to define a domain or a module; one traditional definition is that a domain is a compact subunit of a protein that looks as if it should have independent stability. Some authors refer to a compact unit as a module, and reserve the term domain for a unit that stays together as an evolutionary unit, appearing in partnership with different sets of other domains, or in different orders along the chain. These authors describe the serine protease structure as a single domain comprising two modules.) The giant muscle protein titin contains a long concatenation of up to about 300 modules each of which is homologous to either an immunoglobulin superfamily domain or a fibronectin III domain (Kenny *et al.* 1999). Titin is an extreme example; most modular proteins contain only a few.

Censuses of genomes suggest that many proteins are multimodular. Serres *et al.* (2001) report that of 4401 genes in *E. coli*, 287 correspond to proteins containing 2, 3 or 4 modules. Teichmann *et al.* (2001b, c) have analysed, for enzymes involved in metabolism of small molecules, the distribution and redistribution of domains. The structural patterns of 510 enzymes could be accounted for in total or in part by 213 families of domains. Of the 399 which could be entirely divided into known domains, 68% were single-domain proteins, 24% comprised two domains,

Table 1. *General classification of protein functions (Andrade* et al. *1999)*

Energy
- Biosynthesis of cofactors, amino acids
- Central and intermediary metabolism
- Energy metabolism
- Fatty acids and phospholipids
- Nucleotide biosynthesis
- Transport

Information
- Replication
- Transcription
- Translation

Communication and regulation
- Regulatory functions
- Cell envelope/cell wall
- Cellular processes

and 7% three domains. Only 4 of the 399 had 4, 5 or 6 domains. Teichmann *et al.* (2001b, c) also showed that there are marked preferences for pairing of different families of domains.

Thus multi-domain proteins present particular problems for functional annotation, because domains may possess independent functions, modulate one another's function, or act in concert to provide a single function. However, in some cases the presence of a particular domain or combinations of domains is associated with a specific function. For example, NAD-binding domains appear almost exclusively in dehydrogenases.

4. Classification schemes for protein functions

4.1 General schemes

Several schemes for classification of protein functions have been proposed. We begin with some fairly general categories.

Andrade *et al.* (1999) distinguished the functional classes of proteins involved in energy, information, and communication and regulation. Within these general classes they offered the subdivisions shown in Table 1. These categories comprise fairly general activities rather than individual protein functions. For example, biosynthesis of an amino acid often involves a sequence of reactions catalysed by unrelated enzymes. Despite the differences in the precise function of these enzymes and in their structure and mechanism, all would fall into a single class in this scheme.

Other classifications have appeared in connection with genome sequencing projects. It is interesting to compare an analysis of functional categories suggested for a prokaryotic (*E. coli*) (Table 2) with those suggested for a eukaryote (*Saccharomyces cerevisiae*) (Table 3).

There is a good deal more overlap in these two schemes than first appears. The *E. coli* classes contain a much more precise subdivision of metabolic reactions than the yeast scheme. Perhaps this is an example of the differences in point of view among biochemistry, molecular biology and cell biology. Nevertheless, for purposes of annotating a genome, most people would hope for more specific assignments of function than any of these categories. Note also that the different functions of phosphoglucose isomerase, which is also a neuroleukin, an autocrine motility factor,

Table 2. *Functional groups of proteins for* E. coli (*Blattner* et al. *1997*)

Regulatory function
Putative regulatory proteins
Cell structure
Putative membrane proteins
Putative structural proteins
Phage, transposons, plasmids
Transport and binding proteins
Putative transport proteins
Energy metabolism
DNA replication, recombination, modification, and repair
Transcription, RNA synthesis, metabolism, and modification
Translation, post-translational protein modification
Cell processes (including adaptation, protection)
Biosynthesis of cofactors, prosthetic groups, and carriers
Putative chaperones
Nucleotide biosynthesis and metabolism
Amino acid biosynthesis and metabolism
Fatty acid and phospholipid metabolism
Carbon compound catabolism
Central intermediary metabolism
Putative enzymes
Other known genes (gene product or phenotype known)
Hypothetical, unclassified, unknown

Table 3. *Functional categories suggested for yeast* (*see* http://mips.gsf.de/proj/yeast/catalogues/funcat/)

Metabolism
Energy
Cell cycle and DNA processing
Transcription
Protein synthesis
Protein fate (folding, modification, destination)
Cellular transport and transport mechanisms
Cellular communication/signal transduction mechanism
Cell rescue, defense and virulence
Regulation of/interaction with cellular environment
Cell fate
Transposable elements, viral and plasmid proteins
Control of cellular organization
Subcellular localization
Protein activity regulation
Protein with binding function or cofactor requirement (structural or transport facilitation)
Classification not yet clear cut
Unclassified proteins

and a differentiation and maturation mediator (Jeffery *et al.* 2000) straddle different classes, so that it will be impossible in general to assign individual proteins to unique functional classes.

4.2 The EC classification

The best-known detailed classification of protein functions is that of the EC. Naturally, the EC classification applies only to enzymes. Given our ultimate goal of mapping sequence and

structure onto function, it is important to bear in mind the Commission's emphasis that: '*It is perhaps worth noting, as it has been a matter of long-standing confusion, that enzyme nomenclature is primarily a matter of naming reactions catalysed, not the structures of the proteins that catalyse them.*'

The origin of the EC classification was the action taken by the General Assembly of the International Union of Biochemistry (IUB), in consultation with the International Union of Pure and Applied Chemistry (IUPAC), in 1955, to establish an International Commission on Enzymes. The EC published its classification scheme, first on paper and now on the web (see http://www.chem.qmul.ac.uk/iubmb/enzyme/).

EC numbers (looking suspiciously like IP numbers) contain four fields, corresponding to a four-level hierarchy. For example, EC 1.1.1.1 corresponds to alcohol dehydrogenase, catalysing the general reaction:

an alcohol + NAD = the corresponding aldehyde or ketone + $NADH_2$.

Note that several reactions, involving different alcohols, would share this number; but that the same dehydrogenation of one of these alcohols by an enzyme using the alternative cofactor NADP would be assigned EC 1.1.1.2.

The first number shows to which of the six main divisions (classes) the enzyme belongs:

Class 1. Oxidoreductases
Class 2. Transferases
Class 3. Hydrolases
Class 4. Lyases
Class 5. Isomerases
Class 6. Ligases.

The significance of the second and third numbers depends on the class. For oxidoreductases the second number describes the substrate and the third number the acceptor. For transferases, the second number describes the class of item transferred, and the third number describes either more specifically what they transfer or in some cases the acceptor. For hydrolases, the second number signifies the kind of bond cleaved (e.g. an ester bond) and the third number the molecular context (e.g. a carboxylic ester or a thiol ester). (Proteinases are treated slightly differently, with the third number including the mechanism: serine proteinases, thiol proteinases and acid proteinases are classified separately.) For lyases the second number signifies the kind of bond formed (e.g. C–C or C–O), and the third number the specific molecular context. For isomerases, the second number indicates the type of reaction and the third number the specific class of reaction. For ligases, the second number indicates the type of bond formed and the third number the type of molecule in which it appears. For example, EC 6.1 for C–O bonds (enzymes acylating tRNA), EC 6.2 for C–S bonds (acyl-CoA derivatives), etc. The fourth number gives the specific enzymic activity.

Specialized classifications are available for some families of enzymes; for instance, the MEROPS database by N. D. Rawlings and A. J. Barrett provides a structure-based classification of peptidases and proteinases (see http://www.merops.sanger.ac.uk/).

4.3 Combined classification schemes

Rison *et al.* (2000) have compared functional classifications proposed for genomes. Most are hierarchical, so that the authors could make an attempt to merge them into a 'combined scheme', from which the various classifications could be compared. Of course the different

classifications are not entirely mutually consistent, requiring compromises in integrating them. Their combined scheme is a three-level hierarchy. The top levels are:

(1) metabolism;
(2) process;
(3) transport;
(4) structure and organization of structure;
(5) information pathways;
(6) miscellaneous.

The intermediate and lower levels are increasingly more specific. However, in most cases even the lower level is fairly general; for instance, in the combined scheme of Rison *et al.* (2000), entry 1.3.1 corresponds to metabolism/small molecules/amino-acid metabolism.

Rison *et al.* (2000) map different functional classifications onto their combined scheme and compare coverage. Some gaps are implicit in the design of individual databases. For instance, functions in the general class 'structure' are absent from KEGG – The Kyoto Encyclopaedia of Genes and Genomes (Kanehisa *et al.* 2002) – leaving large gaps in its mapping onto the combined scheme. Some other gaps arise from problems in mapping individual functional classifications onto the combined scheme.

Even this combined scheme does not solve the problem of mapping functions to the level of detail desired for protein annotation. The authors recognize that some of the schemes treated have much higher functional resolution than theirs, but do not integrate that information. They mention but do not treat the EC classification.

Given the goal of mapping a functional classification onto sequence and structure classifications, several problems associated with current functional categorizations are generally recognized. One is that the function is defined without reference to homology in general and structure in particular. The EC, for instance, merges non-homologous enzymes that catalyse similar reactions.

Gerlt & Babbitt (2001), who are among the most thoughtful writers on the subject, pointed out that 'no structurally contextual definitions of enzyme function exist'. They propose a general hierarchical classification of function better integrated with sequence and structure. For enzymes they define:

- *Family.* Homologous enzymes that catalyse the same reaction (same mechanism, same substrate specificity). These can be difficult to detect at the sequence level if the sequence similarity becomes very low.
- *Superfamily.* Homologous enzymes catalysing similar reaction with either (*a*) different specificity or (*b*) different overall reactions with common mechanistic attributes (partial reaction, transition state, intermediate) that share conserved active-site residues.
- *Suprafamilies.* Different reactions with no common feature. Proteins belonging to the same suprafamily would not be expected to be detectable from sequence information alone.

Another problem, that we have already mentioned, is that the traditional biochemist's view of function arises from the study of isolated proteins in dilute solutions, in the presence of carefully controlled concentrations of substrates. The molecular biologist knows that an adequate definition of function must recognize the biological role of a molecule in the living context of a cell (or intracellular compartment) or the complete organism on the one hand, and its role in

a network of metabolic or control processes on the other (Lan *et al.* 2002, 2003). (In addition to the fundamental point of providing a more appropriate definition of function, information about context is often useful in assigning function.) As a result, there is a generic problem with all attempts to force functional classifications into a hierarchical format (see comments of Riley, 1998).

4.4 The Gene Ontology Consortium

A more general approach to the *logical* structure of a functional classification has been adopted by the Gene Ontology Consortium (2000) (see http://www.geneontology.org). Its goal is a systematic attempt to classify function, by creating a dictionary of terms and their relationships for describing molecular functions, biological processes and cellular context of proteins and other gene products. It supports annotation efforts by providing a set of terms that individual annotators or databases may adopt. (By an ontology they mean a set of well-defined terms with well-defined inter-relationships; that is, a dictionary and rules of syntax.)

Organizing concepts of the gene ontology project include the distinctions between:

- *Molecular function.* A function associated which an individual protein or RNA molecule does in itself; either a general description such as 'enzyme', or a specific one such as 'alcohol dehydrogenase'. This is function from the biochemists' point of view.
  and
- *Biological process.* A component of the activities of a living system, mediated by a protein or RNA, possibly in concert with other proteins or RNA molecules; either a general term such as signal transduction, or a particular one such as cyclic AMP synthesis. This is function from the cell's point of view.

Because many processes are dependent on location, gene ontology also tracks:

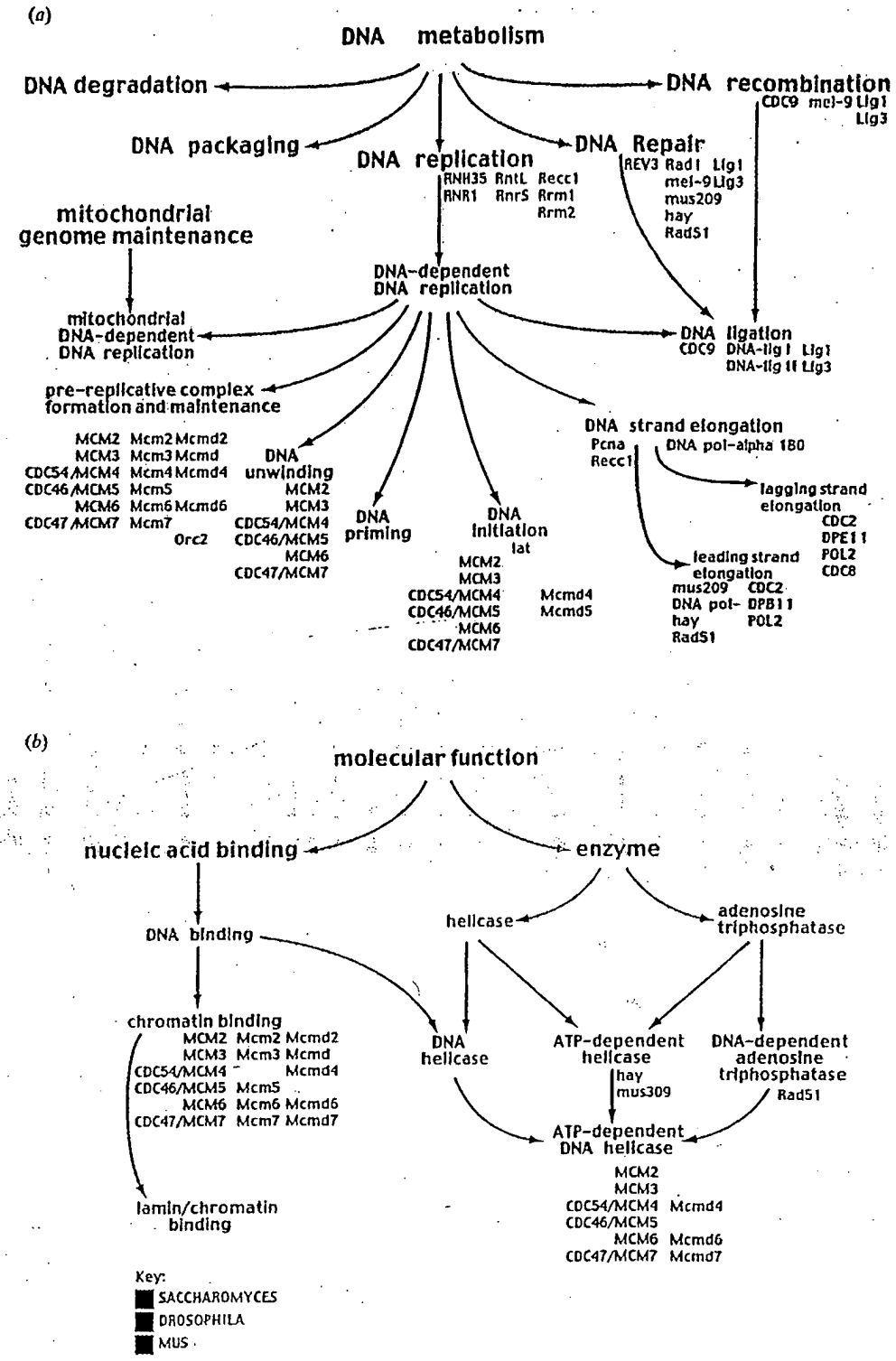
- *Cellular component.* The assignment of site of activity or partners; this can be a general term such as nucleus or a specific one such as ribosome.

An example of the gene ontology classification is shown in Fig. 5. Note that it is more general than a hierarchy. We feel that of the schemes for classification of function that have been proposed, only that of the Gene Ontology Consortium has the possibility of linkage to successful tests of prediction of protein function.

5. Methods for assigning protein function

5.1 Detection of protein homology from sequence, and its application to function assignment

If there is a standard method for predicting protein function, it is the detection of similarity of amino-acid sequence by database searching, and assuming that the molecules identified are homologues with similar functions. Search engines such as PSI-BLAST pull out sequences similar to a query sequence, from general protein sequence databases. The most favourable result is to find that the query sequence is identical or very closely related to that of a well-characterized protein. However, as we have seen, even in these cases the assignment of function may not be correct or complete. The problem of assigning function becomes significantly more complex as the similarity between the unknown sequence and its (putative) homologue falls, except that in



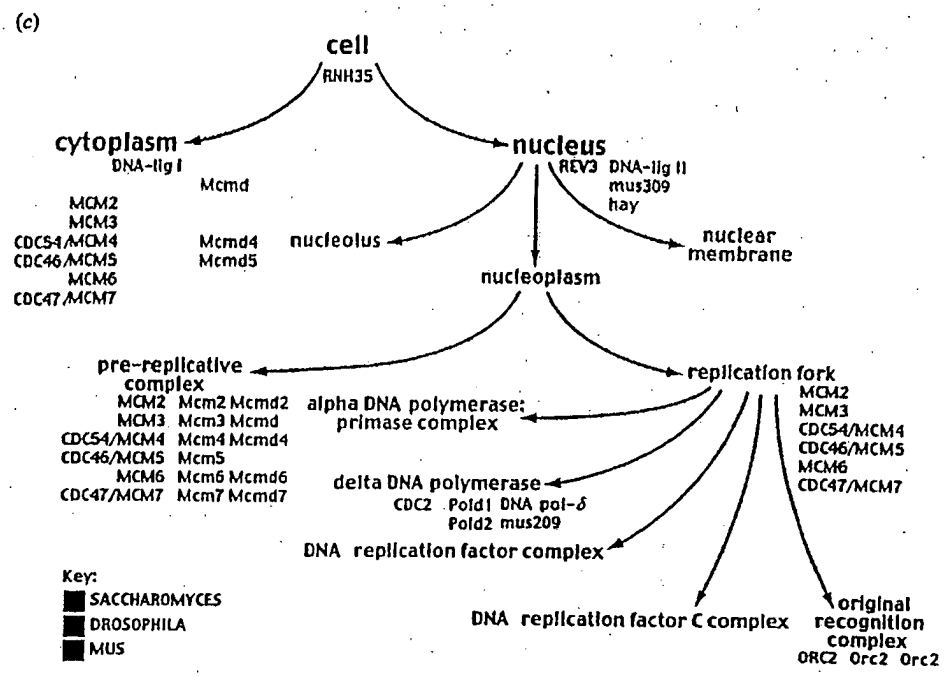


Fig. 5. The Gene Ontology Consortium classification of functions involving DNA metabolism. (From the Gene Ontology Consortium, 2000; reproduced with permission.)

some cases specific sequence signature patterns identify active sites, even in proteins with little overall sequence similarity to homologues of known function. Although the hope is that highly similar proteins will share similar functions, substitution of a single, critically placed amino acid in an active-site residue may be sufficient to alter a protein's role fundamentally.

Several groups have tested correlations between sequence similarity and functional similarity. One senses a feeling, in the relevant scientific community, that can be roughly stated as, 'Yes, we know the collections of horror stories about proteins with very closely related sequences but different functions, but those are rare exceptions, and the inference of function from similarity in sequence works fairly well most of the time.' Does the evidence support this assumption?

Shah & Hunter (1997) determined the sequence similarity of proteins within any EC class. They used a sample of 1327 classes and 15 208 proteins, and tested various similarity thresholds. Their conclusions were that the errors were dominated by false positives, and that it would be better to carry out this kind of analysis at the domain level.

Wilson *et al.* (2000), Todd *et al.* (2001) and Devos & Valencia (2000) reached similar (although not identical) optimistic conclusions.

Wilson *et al.* (2000) conclude that for pairs of single-domain proteins, at levels of sequence identity $\geqslant 40\%$, precise function is conserved, and for levels of sequence identity $\geqslant 25\%$ broad functional class is conserved [according to a functional classification that uses the EC hierarchy for enzymes, and supplements it for material from FLYBASE (Ashburner & Drysdale, 1994) for non-enzymes]. Todd *et al.* (2001) found that for pairs of proteins, both known to be enzymes, slightly $<90\%$ of pairs with sequence identity $\geqslant 40\%$ conserve all four EC numbers. Even at $\geqslant 30\%$ sequence identity they found conservation of three levels of the EC hierarchy for 70% of homologous pairs of enzymes. Devos & Valencia (2000) reached very similar

conclusions; they also reported the ability to predict correctly the agreement of FSSP categories (Holm & Sander, 1999) and SWISS-PROT keywords, as a function of the level of sequence similarity.

Rost (2002), using a wider definition of pairs of sequences identified for comparison – including shorter matching regions – reached more pessimistic conclusions, entirely at variance with those of other investigators. Of pairs of enzymes with >50% sequence identity, he reported that <30% have entirely identical EC numbers. BLAST E values below 10^{-50} were also not sufficient to imply identical function. It should be noted that two pairs of proteins with >50% sequence similarity are expected to have very similar overall structures, <1 Å root-mean-square deviation of over 95% of their backbone atoms, and the active sites may be even more similar in structure (Chothia & Lesk, 1986). Even for pairs of proteins with over 70% residue identity in the optimal alignment (a *very* close relationship indeed), over 30% do not even share the *first* EC number, that is, the general classification! The implication is that to reason successfully from sequence similarity to common function, it is essential to require that the similarity extend over a large enough sector of the sequence, as in the studies of Wilson *et al.* (2000), Todd *et al.* (2001) and Devos & Valencia (2000).

Function prediction from sequence similarity can take advantage of multiple sources of information to back up the prediction from levels of sequence identity alone, and to improve the results in cases of lower sequence similarity than the ~40% identity confidence threshold proposed by Wilson *et al.* (2000), Todd *et al.* (2001) and Devos & Valencia (2000).

Having identified putative homologues, multiple sequence alignments enable identification of conserved residues, the literature may provide crucial information about the family as a whole and the role of conserved residues, and phylogenetic trees can provide information as to whether an unknown protein clusters with a particular functional grouping (Hannenhalli & Russell, 2000; Gu & Vander Velden, 2002). In general, if an unknown protein shares significant sequence similarity with a family of known function, possesses the 'right essential conserved residues' (e.g. active-site residues) then a prediction as to function (proteinase, exonuclease, etc.) can reasonably be proposed. In addition, if the unknown also forms part of a well-supported functional cluster or clade within a phylogenetic tree then a more detailed level of functional prediction may be possible.

Hannenhalli & Russell (2000) examined nucleotidyl cyclases. Changing the specificity between an ATP cyclase and a GTP cyclase requires mutations of only two residues E937K and C1018D. From a common alignment of ATP and GTP cyclases, they were able to identify residues correlated with the change in specificity, including the two crucial positions. Given the sequence of a new enzyme in this family, it could be identified as a family member by overall sequence similarity, and its specificity could be inferred from the residues occupying the selected positions. Hannenhalli & Russell (2000) also showed that a similar analysis permitted prediction of specificity of protein kinases [Motifs were already known that were able to distinguish Ser/Thr from Tyr kinases (Hanks & Hunter, 1995; Hanks *et al.* 1988)].

As a control, an illustration of a negative inference: an evolutionary tree of myotubularin-related proteins permitted Nandurkar *et al.* (2001) to infer that their protein, although related to active phosphatases, lacked the essential catalytic residues and acts as an adapter rather than an enzyme.

Even in the event of a smooth path to successful prediction as outlined above, more questions may be raised than answered. Let us consider an example where we are able to identify an 'unknown' protein as a proteinase through sequence similarity. Immediately the question arises

as to the target of the proteinase (i.e. the physiological substrate), and in addition, what (if any) is its physiological inhibitor(s) or binding partner(s). It may be possible (if a representative structure has been determined) to build molecular models of the unknown proteinase and make basic predictions regarding substrate specificity by examining the nature of the residues lining the predicted S1 subsite. However, this a far cry from being able to predict accurately the *physiological* substrate(s), and thus the biological function. Similar problems exist when attempting to annotate functionally unknown proteins that belong to protein families the primary role of which is to bind other proteins or small molecules – often it is difficult to predict the nature of the binding partner. Thus it appears that relatively straightforward function prediction problems can get bogged down relatively early by questions difficult to answer by common tools of bioinformatics. Nevertheless, even the basic prediction that an unknown protein is a proteinase is valuable information that may guide and accelerate experimental study.

More sensitive database searching engines such as PSI-BLAST and SAM3.0 and other algorithms utilizing profile hidden Markov models (HMMs) allow identification of putative distant homologues. Often these engines are able to detect such similarity in spite of extremely low primary sequence identity (well below the twilight zone – 10–25% sequence identity – into the midnight zone, below 10%). At this level of similarity it is crucial to be able to judge whether or not a match is real and various methods are used to minimize the number of false positives.

Aravind & Koonin (1999) argue that the sequences picked up by sequence similarity sequences represent genuine homologues, on the grounds that current sequence search methods do not pick up even all the proteins known (from structural and other considerations) to be genuine homologues. Of course this is a comment on the state of current sequence searching techniques and the recommended threshold values applied in their use. It may be that more powerful sequence similarity detection programs may in the future pick up sequences that fold into similar structures but are related by convergence rather than homology. The conclusion is that it is important to keep recalibrating the methods in use and – paradoxically – as they grow more powerful, to become more cautious in interpreting their results.

If even close relatives often do not share functions, does the identification of distant putative homologues facilitate functional prediction, or is it a fruitless pursuit? Again, the answer to this question depends on the value placed on a particular threshold of statistical significance in accepting an inference. At best identification of distant similarity to a protein family of known function may suggest a function and allow identification of active-site residues and assignment of the unknown to a general functional class. Even if the relationship is genuine, the unknown may have evolved far from its putative distant homologue. Although it is likely that some general aspects of the mechanism may be common to an unknown and a distant homologue (particularly likely if active-site residues are retained), it is quite possible that fundamental changes in the nature of the substrate may have occurred (e.g. from lipid phosphate to DNA in the case of the AP endonucleases).

Detection of homologues may provide one or more relatives for which the 3D structure is known. This provides another level of information and another test of the prediction. Such a match with a protein of known structure enables a molecular model to be built. Although if the sequence similarity is low the quality of the model may also be low, even an approximate model may allow the compatibility of the unknown sequence with the fold to be assessed (Schonbrun *et al.* 2002). Furthermore, because the active sites of enzymes often comprise the most highly conserved and structurally similar regions it may be possible to build a surprisingly detailed model around the active site, even if overall sequence similarity is low. The two examples given

in the introduction from the structural genomics of *Haemophilus influenzae* illustrate the experience that this approach sometimes works and sometimes does not.

Even if the results of an experimental structure determination are not available, theoretical methods of structure prediction may be useful in identifying putative remote homology (Schonbrun *et al.* 2002; Tramontano, 2003; Kinch *et al.* In Press).

The situation for multi-domain proteins is even more complex. Although it may be relatively straightforward to predict the role of some of the domains using the methods described, others may prove more challenging. Thus a complete functional description of a multi-domain protein of unknown function may be limited if it contains one or more domains that cannot be accurately annotated. Furthermore the possibility of domains acting in concert with one another to modulate the behaviour of the complete molecule is difficult to predict.

5.2 Detection of structural similarity, protein structure classifications, and structure/function correlations

It is well known that structure changes more conservatively than sequence during evolution. There are many cases of distantly related homologues assignable from shared structures with no recognizable relationship between the sequences. The 3D analogue of sequence alignment is alignment by structural analogy: establishment of correspondences between pairs of residues that occupy the same geometric positions in two protein structures. Many algorithms have been implemented for this task (reviewed by Koehl, 2001).

DALI (Holm & Sander, 1993) is based on the observation that inter-residue contact patterns are among the best preserved features of protein structures (Lesk & Chothia, 1980). The DALI web server (see http://www.ebi.ac.uk/dali/) will screen a novel protein structure against the Protein Data Bank and report the most similar structures and the alignment of the sequences. DALI is used routinely by X-ray crystallographers and NMR spectroscopists to provide a preliminary classification of each new structure.

Several authors have applied the known structures to infer homology among proteins too distantly related to be identified as homologues from the sequences alone. They have created databases merging structures, sequences and the greater reliability of homology detection and alignment attainable by use of structural information (Holm & Sander, 1999; Przytycka *et al.* 1999; Aloy *et al.* 2002).

A hierarchical structural classification of protein domains of known structure, based on the DALI program, is available on the web (Holm & Sander, 1999). Two other major databases of classifications of protein structures are the Structural Classification of Proteins (SCOP) (Murzin *et al.* 1995; Lo Conte *et al.* 2002) and CATH (Pearl *et al.* 2003). There are many others, tabulated in Ouzounis *et al.* (2003). SCOP depends crucially on manual curation by A. G. Murzin. CATH is based on a structural-alignment program, SSAP (Taylor & Orengo, 1989). Most classification schemes for sequences and structures are expressed as hierarchical clusterings. The most similar items are grouped together at the lowest level. The sets of linked items are progressively merged to form successive levels of the hierarchy. For instance, the SCOP database has as its basis individual domains of proteins. Sets of domains are grouped into *families* of homologues, for which the similarities in structure, sequence, and sometimes function imply a common evolutionary origin. Families containing proteins of similar structure and function, but for which the evidence for evolutionary relationship is suggestive but not compelling, form *superfamilies*. Superfamilies that share a common folding topology, for at least a large central portion of

the structure, are grouped as *folds*. Finally, each fold group falls into one of the general *classes*. The major classes in SCOP are α, β, $\alpha+\beta$, α/β, multi-domain proteins, membrane and cell surface proteins, and miscellaneous small proteins, which often have little secondary structure and are held together by disulphide bridges or ligands.

Several groups have attempted to correlate protein structure and function (Hegyi & Gerstein, 1999; Thornton *et al.* 1999). Hegyi & Gerstein (1999) correlated the enzymes in the yeast genome between their fold classification in SCOP (Lo Conte *et al.* 2002) and their EC functional categories, via the annotations in SWISS-PROT. They identified 8937 single-domain proteins that could be assigned both a fold and a function.

The broadest categories of structure were from the top of the SCOP hierarchy, including the all-α, all-β, α/β, $\alpha+\beta$, multi-domain, and small classes. The broadest categories of function were from the top of the EC hierarchy: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases; plus an additional category, non-enzymes. There are therefore 6 (structural classes) $\times$ 7 (functional classes) = 42 possible combinations of highest-level correlates. By using finer classifications of structure and function (down to the third level of EC numbers) there are a total of 21 068 potential fold-function combinations. Only 331 of these are observed, among the 8937 proteins analysed.

The observed distribution is highly non-random. Non-enzymic functions account for 59% of the sequences of which well over half are in the all-α or all-β fold category. Of the enzymes, the most popular combinations were α/β folds among oxidoreductases and transferases, and all-β and $\alpha+\beta$ hydrolases.

Knowing the structure of a domain, what can be inferred about its function? Many folds are compatible with very different activities. The five most 'versatile' folds are the TIM barrel, α–β hydrolase, the NAD-binding fold, the P-loop-containing NTP hydrolase fold, and the ferredoxin fold. Conversely, the functions carried out by the most different types of structure are glycosidases and carboxylases. These two functions are carried out by seven different fold types, from three different fold classes.

What we are looking for, however, are cases where structure provides reliable clues to function. In their cross table, Hegyi & Gerstein (1999) show several folds that appear in combination with only one function. These appear to have predictive significance for function. Of course one cannot tell whether this is just because they are rare folds, and whether the correlation will hold up as the databases grow.

5.3 Function prediction from amino-acid sequence

Despite the progress in structural genomics projects, most proteins encoded in newly sequenced genomes are known from their amino-acid sequences alone. A major problem in genome annotation is that of assigning their functions. Note that not only are the 3D structures in most cases unknown, there is generally no information even about cofactors or post-translational modifications, which are often essential for function.

There are two basic approaches to prediction of protein function from amino-acid sequence alone, focused on (1) overall sequence similarity and (2) signature patterns of active sites, or motifs (Bork & Koonin, 1996). We have already discussed the standard method based on the assumption that in at least many cases evolutionary divergence is slow enough to permit recognition of homologues that may have the same or at least similar structures and functions. Often the general similarity of sequences reflects a similarity in overall folding pattern, and particular

residues within the fold may form a localized active site. Clearly the conservation of active-site residues is important in reasoning from sequence similarity to functional similarity. Indeed, in some cases it is possible to cut short the reasoning and to recognize the residues comprising the active site from a specific signature pattern or motif within the sequence. However, although many motifs do reflect functional active sites, others reflect positions for post-translational modification (e.g. glycosylation sites), or structural signals (e.g. N and C caps of α-helices), or signal sequences, with no direct functional implications.

Attwood (2000) has described general methods for deducing sequence patterns. All start with (or produce) a multiple sequence alignment, and seek to identify common distinctive features of particular positions of the sequence. These features may involve:

(1) A motif describing a single consecutive set of residues.
(2) Multiple motifs – a combination of several motifs involving separate consecutive sets of residues.
(3) Profile methods, based on entire sequences and weighting different residue positions according to the variability of their contents. Extensions and generalizations of profile methods, including HMMs, are among the most sensitive detectors of distant homology based entirely on sequence data that we have.

5.3.1 Databases of single motifs

Motifs may be expressed in terms of uniquely defined sequences, such as

GWTLNSAGYLLGP,

which characterizes the neuropeptide galanin. Or, motifs may contain alternative residues; for instance [LIVMF]-T-T-P-P-[FY], the signature of N-4 cytosine-specific DNA methylases. Here [LIVMF] means that that first position may contain *any* of the amino acids L, I, V, M, or F, followed by the unique sequence TTPP, followed by a position that may contain either F or Y. It is easy to indicate a site which excludes a specific amino acid by bracketing the other 19, or by using the notation {P} to indicate 'any amino acid except proline'. Motifs can contain 'wild cards' (which permit any of the 20 amino acids at a position) and 'spacers'; for instance, L-x(6)-L-x(6)-L-x(6)-L, the signature pattern of the leucine zipper which appears in some eukaryotic transcription regulator proteins. The pattern specifies four leucines each separated by six residues each of which may be any amino acid. More generally, a signature pattern may be specified by a 'regular expression', which allows for a wider range of alternative patterns and variable distances between residue positions. It is simplest to search for exact matches to the patterns, but algorithms that allow for some mismatches are available (see e.g. Gusfield, 1997; Crochemore & Rytter, 2003).

Attempts to apply data mining techniques to pattern discovery in biological sequences are by now a heavy industry with an enormous literature (see e.g. Floratos *et al.* 2001).

One very important set of results of this kind of work, PROSITE (Sigrist *et al.* 2002) contains a collection of motifs covering a wide range of groups of proteins, together with retrieval software to check a submitted sequence for the presence of one or more motifs. The motifs are calibrated to indicate the number of false negatives and positives to be expected. The [LIVMF]-T-T-P-P-[FY] motif detects all N-4 cytosine-specific DNA methylases, but also picks up false positives. The L-x(6)-L-x(6)-L-x(6)-L motif is least specific, missing one known leucine zipper (L-myc, which contains a methionine instead of one of the leucines)

and promiscuously picking up hundreds of other sequences from many different types of proteins.

Thornton *et al.* (1999) have investigated the structural implications of conserved sequence motifs. Typically these are involved in conserved substructures contributing to a common function. Kasuya & Thornton (1999) have confirmed that PROSITE motifs reflect common 3D structural patterns by analysis of protein structures in which they appear. Kasuya & Thornton (1999) found examples among proteins of known structure of 553 of the 1265 PROSITE patterns available at the time of their work. In most cases the residues matching a given PROSITE pattern in different proteins had similar 3D structures as measured by the root-mean-square deviation of the Cα atoms. Some of the exceptions observed are biologically interesting. For instance, among the matches to the 12-residue TRYPSIN_SER pattern that includes the active-site serine of the trypsin family of serine proteinases

[DNSTAGC]-[GSTAPIMVQH]-x(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]
-[LIVMFYSTANQH]

outliers in conformation space included proenzymes, for which it is known that the region matching the pattern undergoes conformational change upon activation.

Todd *et al.* (2001, 2002) have collected cases of homologous enzymes, some but not all of which catalyse the same reaction, in which residues equivalent in their contribution to catalysis appear at non-equivalent positions in the active site. Examples include human alcohol dehydrogenases in classes 1β and IIIχ, which have 62% overall residue identity in their sequence alignment, but in which the active site Thr and His appear in different sequence patterns: 48Thr-49Asp-50Asp-51His or 47His-48Thr; and β-lactamases in classes A and C, in which the catalytic residues appear on different structural elements.

Several authors have sought to extend motif searching to three dimensions. Given that motifs tend to correspond to regions of conserved structure linked to function, Wallace *et al.* (1996) searched known protein structures for the Ser-His-Asp catalytic triad of trypsin-like serine proteinases. The identified all known serine proteinases in their dataset, plus triglycerol lipases which share the catalytic triad.

de Rinaldis *et al.* (1998) derived 3D profiles from a single protein structure or a set of aligned structures. They applied their results to identifying proteins with matching surface patches. Analysis of the 3D profiles of ATP and GTP binding P-loop proteins identified a positively charged phosphate-binding residue (Arg or Lys) in a position conserved in space but not in sequence.

In a similar approach, Jackson & Russell (2000, 2001) have identified regions with conformations similar to those of PROSITE motifs, but not necessarily sharing sequence similarity with them. They were able to identify serine proteinase inhibitors that contain regions similar in conformation to the loops in known inhibitors that have a common structure that docks to the proteinase.

5.3.2 Databases of profiles

Given a multiple sequence alignment, it is usually the case that some positions show high variability while others show high conservation. To detect other sequences that share the pattern, a weighted alignment of a target sequence with the alignment table can be carried out, giving

higher weight to matches at highly conserved positions. Profiles could alternatively be based on the local regions of high conservation that went into motifs.

Associated with PROSITE is a compendium of profiles characterizing entire domains. Because the matching of such profiles is sensitive to the sequences of entire domains, it is less likely to return false positives; but because the information contained in the most conserved part of the sequences is eroded, it may lose sensitivity relative to motif matching.

An alternative approach to describing a set of homologous sequences is HMMs (Eddy, 1996). HMMs represent successive positions in a probabilistic way. They are more general than simple profiles, and do a better job of discriminating homologues from non-homologues, provided that they are trained with correct alignments. HMMs currently provide the most sensitive methods for detecting distant homologues given only the amino-acid sequence of a query protein.

Pfam is a database of multiple alignments of protein domains, and the HMMs built from them (Bateman *et al.* 2002). Search software permits detection of whether a query sequence belongs to any of the families in Pfam.

The Superfamily database is a library of HMMs for all proteins of known structure (Gough *et al.* 2001). Its goal is to identify, from protein sequences, domains with folds corresponding to one or more known structures.

5.3.3 Databases of multiple motifs

We have pointed out that motifs may be more specific than profiles because they focus on well-conserved active sites. But a weakness of single-motif patterns is that an active site of a protein may be defined by regions that are distant in the sequence although nearby in space. Single-motif patterns are also necessarily based on characteristics of single domains, whereas it may be useful to identify proteins by the presence of more than one domain. Multiple-motif databases aim to remedy these problems.

BLOCKS (Henikoff *et al.* 2000) and PRINTS (Attwood, 2002; Attwood *et al.* 2002a, 2003) are databases of multiple motifs, typically ~20 residues long, presented in the form of ungapped multiple sequence alignments. PRINTS but not BLOCKS contains biological documentation of the significance of the motifs. Search software can identify matches to individual motifs in a query sequence. There is flexibility to define how many of the motifs match, to what stringency, to define a 'hit'.

If different sets of sequences match different individual motifs one has the additional possibility of classifying subsets of a family of homologues, and inferring evolutionary trees. For instance, Attwood (2001, 2002) has used motifs to classify the important family of G protein-coupled receptors (GPCRs), a large family of cell-surface proteins that detect and signal hormones and growth factors, and mediate the senses of sight and smell. Particular motivation for classifying subtypes is the fact the GPCRs are common drug targets. Potential for improvements in specificity would have important clinical consequences.

The PRINTS database contains a seven-motif fingerprint for GPCRs – each motif corresponding to one of the transmembrane helices. Additional sets of motifs identify subfamilies of GPCRs and receptor subtypes. Some but not all of these motifs overlap the general family fingerprint. Mapping of the motifs onto the structure of rhodopsin shows what structural features distinguish the subclasses (Attwood *et al.* 2002b).

To apply these databases to prediction of protein function, it should be kept in mind that profiles or HMMs are sensitive to overall folding pattern, sometimes at the expense of focus on

Table 4. *Some of the databases of protein family classifications*

Database	Contents	Reference
Primarily sequence based		
BLOCKS+	Families	Henikoff *et al.* (2000)
COG	Families	Tatusov *et al.* (2001)
HSSP	Protein families including proteins	Holm & Sander (1999)
InterPro	Families/domains	Mulder *et al.* (2003)
Pfam	HMM-based families	Bateman *et al.* (2002)
PIMA	Domains	
PIR-ALN	Domains, families, superfamiles	Srinivasarao (1999)
PRINTS	Families	Attwood (2002)
iProClass	Domains, families, superfamilies	Huang *et al.* (2003)
ProDom	Domains	Servant *et al.* (2002)
PROSITE	Families	Sigrist *et al.* (2002)
ProtoMap	Families	Yona *et al.* (2000)
PROT-FAM	Domains, families, superfamilies	Mewes *et al.* (1997)
SBASE	Domains of known structure	Vlahovicek *et al.* (2002)
Hierarchical protein structure classifications		
SCOP	Domains	Lo Conte *et al.* (2002)
CATH	Domains	Orengo *et al.* (2002)
DALI domain dictionary	Domains	Dietmann & Holm (2001)

specific active-site residues. Conversely, some motifs are sensitive to active-site residues but in their insensitivity to features of the sequence as a whole may pick up non-homologous proteins as false positives.

Among these classes of method, a combination of a profile and motif match would therefore seem to be the most reliable criterion for function assignment (see Chen & Jeong, 2000).

5.3.4 Precompiled families

Several groups have applied tools for sequence matching to full sequence databases, or used structural similarity, to classify proteins (Table 4). Note that the exact definitions of the categories vary among the databases.

InterPro is an umbrella database that attempts to integrate the contents, features, and annotation of several individual databases of protein families, domains, and functional sites (Mulder *et al.* 2003). It subsumes, but is not limited to, information from PROSITE, Pfam, PRINTS, SMART and ProDom databases, and contains links to others including the Gene Ontology Consortium functional classification. It intends to assimilate additional databases, including structural databases. Resistance is futile.

An InterPro entry is a description of a protein family, domain, repeat, or site of post-translational modification, and links to other databanks, and original literature. Annotations from the source databases are merged. Each entry includes links to relevant terms from the Gene Ontology Consortium classification schemes.

5.3.5 Function identification from sequence by feature extraction

Although information about function must be contained implicitly in amino-acid sequences, it is obscure. It can be seen that, even using structure as an intermediate stepping-stone between

sequence and function does not satisfactorily resolve the problem. Brunak and his colleagues have examined an alternative intermediate between sequence and function (Jensen *et al.* 2002). They reasoned that information about function should be contained in a spectrum of features of proteins, including secondary structure, post-translational modifications, protein sorting, and general properties of the amino-acid composition such as the isoelectric point. Using neural networks they predicted the following features from protein sequences, and correlated the results with functional classes:

- extinction coefficient;
- grand average hydrophobicity;
- number of negative residues;
- number of positive residues;
- O-glycosylation;
- serine/threonine phosphorylation;
- tyrosine phosphorylation;
- N-glycosylation;
- PEST-rich regions;
- secondary structure;
- subcellular location;
- low complexity regions;
- signal peptides;
- transmembrane helices.

They recognized that the predictions of the features would be imperfect, but this need not fatally degrade their prediction of function.

The combined networks were trained to recognize a general set of functional classes based on categories originally defined by Riley (1993), and, within the proteins predicted to be enzymes, the EC classification. As a measure of the quality of the results, for the general categories, at a level of thresholding giving 70% correct predictions, the range of false positives varied from below 10% to below 40%, with most categories giving about 20% false positives. (A sensitivity of 70 with 20% false positives means that if a large number of novel sequences are submitted to the procedure, and this set of sequences contains 100 examples of proteins in some functional class, the network will report that 90 of the proteins are in that functional class; 70 of the predictions will be correct and 20 will correspond to proteins outside the functional class.)

By analysing the networks, Jensen *et al.* (2002) were also able to analyse which particular combinations of features were the most effective signals for specific functional types.

5.4 Methods making use of structural data

Several groups have developed methods to apply structural information, in most cases in combination with sequence information, to interpret function.

Shapiro & Harris (2000) and Teichmann *et al.* (2001a) illustrate the power of structure, including but not limited to identifying distant relationships not derivable from sequence comparisons.

(1) Identification of structural relationships unanticipated from sequence can suggest similarity of function. The crystal structure of AdipoQ, a protein secreted from adipocytes, showed

a similarity of folding pattern to that of tumour necrosis factor. The inference that AdipoQ is a cell-signalling protein was subsequently verified.

(2) The histidine triad proteins are a broad family with no known function. Analysis of their structures indicated a catalytic centre and nucleotide-binding site, identifying them as a nucleotide hydrolase. Note that this did *not* depend on detection of a distant homology.

(3) Structural similarity of a gene product of unknown function from *Methanococcus jannischii* and other proteins containing nucleotide-binding domains led to experiments showing it to be a xanthine or inosine triphosphatase (Hwang *et al.* 1999).

Like most sequence-based methods, these structure-based methods proceed by searching for homologues. It is well known that distant homology is frequently more easily detectable in structure than in sequence. However, one must recognize that the more distant the relationship, the less reliable the inference of common function. In general, structure does not permit unambiguous assignment of a precise function, but can provide guidance to experiments that can do so.

Several groups have attempted to determine the common functionally active site of a family of proteins. Lichtarge *et al.* (1996a) have developed an evolutionary trace method to define binding surfaces common to protein families. They extract functionally important residues from sequence conservation patterns and map them onto the protein surface to identify functional clusters.

Given a set of homologous sequences, and at least one structure, the goal of the evolutionary trace method is to identify surface sites implicated in function. The assumptions of the method are:

(1) The set of proteins has a common surface-exposed active site.

(2) The homologous sequences produce similar structures, that retain the location in molecular space of the active site.

(3) The functional site is less subject to mutation that average surface sites.

(4) Those mutations in the functional site that do occur are not random but create discrete sets of structures with shifts in function (see also Golding & Dean, 1998; Gu, 1999).

The method begins by forming a multiple sequence alignment, from which the molecules are hierarchically clustered into a tree. By choosing different levels in the hierarchy, clusters of different size may be extracted. If different functions are known in the family, the clusters are chosen to reflect subgroups with different function. By choosing larger or smaller clusters, grosser or finer resolution in function distinction may be made. For each cluster in the partition, form a consensus sequence alignment. Then co-align all the consensus sequences. The residues can be divided into (*a*) those that are absolutely conserved, (*b*) those that are conserved within clusters but differ between clusters, and (*c*) unconserved positions. By mapping the conserved residues onto the structure, a pattern is observed that defines a surface patch predicted to correspond to the active site.

Lichtarge *et al.* (1996a) applied their method to SH2 and SH3 signalling domains, and the DNA-binding domain of nuclear hormone receptors. Their results correctly identified the known functional sites in these molecules. If the evolutionary trace method depended on a classification induced by known functional divergence, as in these test cases, it would be arguable that it was really a method for assigning structure to function rather than function to structure. However, it can be applied using trees from other sources, and the classifications they induce; for instance, those based solely on multiple sequence alignments.

Successful predictions by the evolutionary trace method include identification of the functional surface in families of G protein α-subunits (Lichtarge *et al.* 1996b) and regulators of G protein signalling (Sowa *et al.* 2000, 2001). Both cases were *blind* predictions subsequently verified by experiment. The success of the evolutionary trace method has led to its being taken up and developed by a number of groups (Aloy *et al.* 2001; Lichtarge & Sowa, 2002; Madabushi *et al.* 2002; Yao *et al.* 2003).

Irving *et al.* (2001) applied the idea that active sites tend to be among the structurally best conserved parts of a protein, by using superposition methods to extract regions of the lowest root-mean-square deviation of Cα atoms in a pair of proteins of known structure. They tested the method on a pair of proteins – YabJ from *B. subtilis* (PDB entry 1qd9) and YjgF from *E. coli* (1qu9) – related to chorismate mutase. Without using any information from chorismate mutase, their program suggested that YabJ and YjgF share an active site, which occupies a similar region of their structures as the active site of chorismate mutase.

It should be emphasized that identification of an active site is not *per se* an identification of function, but an important step towards one. Once a binding site is targeted, the identification of a ligand is, computationally, the same problem faced in drug design, for which a great deal of mature algorithms and software exist (Finn & Kavraki, 1999).

Moreover, the mode of binding of a ligand does not always correlate with sequence or structural similarity. Cappello *et al.* (2002) studied the mode of binding of the adenine ring in different proteins. Their conclusion was that proteins with similar folds can bind adenine in different ways, and (interesting but less relevant for possible methods for function prediction) proteins with dissimilar structures and functions can bind adenine in similar ways.

6. Applications of full-organism information: inferences from genomic context and protein interaction patterns

For proteins encoded in complete genomes, approaches to function prediction making use of contextual information and intergenomic comparisons are useful (Marcotte *et al.* 1999; Huynen *et al.* 2000; Huynen & Snel, 2000; Kolesov *et al.* 2001, 2002).

(1) *Gene fusion.* A composite gene in one genome may correspond to separate genes in other genomes. The implication is that there is a relationship between the functions of these genes.
(2) *Local gene context.* It makes sense to co-regulate and co-transcribe components of a pathway. In bacteria, genes in a single operon are usually functionally linked.
(3) *Interaction patterns.* As part of the development of full-organism methods of investigation, data are becoming available on patterns of protein interactions (Xenarios *et al.* 2002). The network of interactions reveals the function of a protein.
(4) *Phylogenetic profiles.* Pellegrini *et al.* (1999) have exploited the idea that proteins in a common structural complex or pathway are functionally linked and expected to co-evolve. For each protein encoded in a known genome, they construct a phylogenetic profile that indicates which organisms contain a homologue of the protein in question. Clustering the profiles identifies sets of proteins that co-occur in the same group of organisms. Some relationship between their functions is expected.

For instance, *E. coli* ribosomal protein RL7 has homologues in 10 out of 11 eubacterial genomes, but no homologue appears in an archaeal genome (Pellegrini *et al.* 1999). Most of the *E. coli* proteins that share the phylogenetic profile of RL7 have ribosome-associated functions.

If the function of RL7 were unknown one could infer that it is associated in some way with the ribosome. Comparison of keywords in SWISS-PROT annotations affords a general test of this approach. Of sets of non-homologous proteins with similar phylogenetic profiles had, on average, 18% of SWISS-PROT keywords in common.

There need be no sequence or structural similarity between the proteins that share a phylogenetic distribution pattern. One unusual and very welcome feature of this method is that it is one of the few that derives information about the function of a protein from its relationship to *non-homologous* proteins (Marcotte *et al.* 1999; Pellegrini *et al.* 1999).

7. Conclusions

The problem of prediction of function from amino-acid sequence and protein structure is far from being satisfactorily solved.

Some problems are hard only because they are difficult; others are hard because they are both difficult and messy. The prediction of protein structure from amino-acid sequence is difficult, but we know that nature has an algorithm and all we have to do is find it, and given any procedure we can easily decide whether the answer is correct or not. The prediction of protein function is messy, partly because function is a fuzzy and multi-faceted concept, and partly because very small (or even no) changes in amino-acid sequence are compatible with large changes in function.

It appears that the most general classification of function is that produced by the Gene Ontology Consortium. Their results have the advantage of being appropriate to both biochemistry and biology, at the expense of greater logical complexity.

Many of the methods that have been applied to function prediction work part of the time but none is perfect. Moreover, the more expert the analysis of the results applied, the better the predictions are. This makes it difficult to envisage a purely 'black-box' automatic annotation machine for new whole-genome sequences. In most cases, predictions suggest, but do not determine, the general class of function. Their most useful effect is to guide investigations in the laboratory to confirm, or refute, the prediction, and, even if correct, to define the function in greater detail.

We conclude that predictions are useful but no substitute for work in the laboratory. Indications from theory may indict, but only experimental evidence can convict.

8. Acknowledgements

A.M.L. is supported by The Wellcome Trust and J.C.W. by the National Health and Medical Research Council of Australia and the Australian Research Council. We thank Drs T. K. Attwood, M. Bashton, P. Bork, D. S. Eisenberg, M. Helmer-Citterich, O. Herzberg, E. V. Koonin, O. Lichtarge, L. Lo Conte, J. Moult, B. Rost and A. Valencia, for helpful suggestions.

9. References

Aloy, P., Oliva, B., Querol, E., Aviles, F. X. & Russell, R. B. (2002). Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. *Protein Science* 11, 1101–1116.

Aloy, P., Querol, E., Aviles, F. X. & Sternberg, M. J. (2001). Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *Journal of Molecular Biology* 311, 395–408.

Anantharaman, V., Aravind, L. & Koonin, E. V. (2003). Emergence of diverse biochemical activities in

evolutionary conserved structural scaffolds of proteins. *Current Opinion in Structural Biology* 7, 12–20.

Andrade, M. A., Ouzounis, C., Sander, C., Tamames, J. & Valencia, A. (1999). Functional classes in the three domains of life. *Journal of Molecular Evolution* 49, 551–557.

Aravind, L. & Koonin, E. V. (1999). Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches. *Journal of Molecular Biology* 287, 1023–1040.

Ashburner, M. & Drysdale, R. (1994). Flybase: the *Drosophila* genetic database. *Development* 120, 2077–2079.

Attwood, T. K. (2000). The quest to deduce protein function from sequence: the role of pattern databases. *International Journal of Biochemistry and Cell Biology* 32, 139–155.

Attwood, T. K. (2001). A compendium of specific motifs for diagnosing GPCR subtypes. *Trends in Pharmacological Science* 22, 162–165.

Attwood, T. K. (2002). The PRINTS database: a resource for identification of protein families. *Briefings in Bioinformatics* 3, 252–263.

Attwood, T. K., Blythe, M., Flower, D. R., Gaulton, A., Mabey, J. E., Maudling, N., McGregor, L., Mitchell, A., Moulton, G., Paine, K. & Scordis, P. (2002a). PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research* 30, 239–241.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N. & Mitchell, A. L. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research* 31, 400–402.

Attwood, T. K., Croning, M. D. & Gaulton, A. (2002b). Deriving structural and functional insights from a ligand-based hierarchical classification of G protein-coupled receptors. *Protein Engineering* 15, 7–12.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. L. (2002). The Pfam protein families database. *Nucleic Acids Research* 30, 276–280.

Blattner, F. R., Plunkett, G. R., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. & Yuan, Y. (1998). Predicting function: from genes to genomes and back. *Journal of Molecular Biology* 283, 707–725.

Bork, P. & Koonin, E. V. (1996). Protein sequence motifs. *Current Opinion in Structural Biology* 6, 366–376.

Bork, P. & Koonin, E. V. (1998). Predicting functions from protein sequences – where are the bottlenecks? *Nature Genetics* 18, 313–318.

Brenner, S. E. (1999). Errors in genome annotation. *Trends in Genetics* 15, 132–133.

Brenner, S. E. (2001). A tour of structural genomics. *Nature Reviews Genetics* 2, 801–809.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nature Genetics* 23, 151–157.

Cappello, V., Tramontano, A. & Koch, U. (2002). Classification of protein based on the properties of the ligand-binding site: the case of adenine-binding proteins. *Proteins: Structure, Function and Genetics* 47, 106–115.

Chance, M. R., Bresnick, A. R., Burley, S. K., Jiang, J. S., Lima, C. D., Sali, A., Almo, S. C., Bonanno, J. B., Buglino, J. A., Boulton, S., *et al.* (2002). Structural genomics: a pipeline for providing structures for the biologist. *Protein Science* 11, 723–738.

Chen, R. & Jeong, S. S. (2000). Functional prediction: identification of protein orthologs and paralogs. *Protein Science* 9, 2344–2353.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal* 5, 823–826.

CODATA Task Group on Biological Macromolecules and colleagues (2000). Quality control in databanks for molecular biology. *BioEssays* 22, 1024–1034.

Cohen, F. E. & Prusiner, S. B. (1998). Pathologic conformations of prion proteins. *Annual Review of Biochemistry* 67, 793–819.

Copley, R. R. & Bork, P. (2000). Homology among (beta-alpha)(8) barrels: implications for the evolution of metabolic pathways. *Journal of Molecular Biology* 303, 627–641.

Crochemore, M. & Rytter, W. (2003). *Jewels of Stringology*. London: World Scientific.

de Rinaldis, M., Ausiello, G., Cesareni, G. & Helmer-Citterich, M. (1998). Three-dimensional profiles: a new tool to identify protein surface similarities. *Journal of Molecular Biology* 284, 1211–1221.

Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Structure, Function and Genetics* 41, 98–107.

Devos, D. & Valencia, A. (2001). Intrinsic errors in genome annotation. *Trends in Genetics* 17, 429–431.

Doerks, T., Bairoch, A. & Bork, P. (1998). Protein annotation: detective work for function prediction. *Trends in Genetics* 14, 248–250.

Dietmann, S. & Holm, L. (2001). Identification of homology in protein structure classification. *Nature Structural Biology* 8, 953–957.

Doolittle, R. F. (1994). Convergent evolution: the need to be explicit. *Trends in Biochemical Science* 19, 15–18.

Eddy, S. R. (1996). Hidden Markov models. *Current Opinion in Structural Biology* 6, 361–365.

Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. (2000). Protein function in the post-genomic era. *Nature* 405, 823–826.

Eisenstein, E., Gilliland, G. L., Herzberg, O., Moult, J., Orban, J., Poljak, R. J., Banerjei, L., Richardson, D. & Howard, A. J. (2000). Biological function made crystal clear – annotation of hypothetical proteins via structural genomics. *Current Opinion in Biotechnology* **11**, 25–30.

Finn, P. W. & Kavraki, L. E. (1999). Computational approaches to drug design. *Algorithmica* **25**, 347–371.

Floratos, A., Rigoutsos, I., Parida, L. & Gao, Y. (2001). DELPHI: A pattern-based method for detecting sequence similarity. *IBM Journal of Research and Development* **45**, 455–474.

Galperin, M. Y. & Koonin, E. V. (2002). *Sequence–Evolution–Function: Computational Approaches in Comparative Genomics*. Dordrecht: Kluwer Academic Publishers.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Research* **8**, 779–790.

Ganfornina, M. D. & Sánchez, D. (1999). Generation of evolutionary novelty by functional shift. *BioEssays* **21**, 432–439.

Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–28.

Gerlt, J. A. & Babbitt, P. C. (2000). Can sequence determine function? *Genome Biology* **1**, reviews0005.

Gerlt, J. A. & Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annual Review of Biochemistry* **70**, 209–246.

Gettins, P. G. (2002). Serpin structure, mechanism, and function. *Chemical Review* **102**, 4751–4804.

Gilliland, G. L., Tereshkov, A., Obmolova, G., Tordova, M., Thanki, N., Ladner, J., Herzberg, O., Lim, K., Zhang, H., Huang, K., *et al.* (2002). Assisting functional assignment for hypothetical *Haemophilus influenzae* gene products through structural genomics. *Current Drug Targets Infectious Disorders* **2**, 339–353.

Golding, G. B. & Dean, A. M. (1998). The structural basis of molecular adaptation. *Molecular and Biological Evolution* **15**, 355–369.

Gough, J., Karplus, K., Hughey, R. & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology* **313**, 903–919.

Grishin, N. (2001). Fold change in evolution of protein structures. *Journal of Structural Biology* **134**, 167–185.

Gu, X. (1999). Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution* **16**, 1664–1674.

Gu, X. & Vander Velden, K. (2002). DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* **18**, 500–501.

Gusfield, D. (1997). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Hanks, S. K. & Hunter, T. (1995). Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB Journal* **9**, 576–596.

Hanks, S. K., Quinn, A. M. & Hunter, T. (1988). The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science* **241**, 42–52.

Hannenhalli, S. S. & Russell, R. B. (2000). Analysis and prediction of functional sub-types from protein sequence alignments. *Journal of Molecular Biology* **303**, 61–76.

Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W., Kenyon, G. L., Babbitt, P. C., Gerlt, J. A., Petsko, G. A. & Ringe, D. (1998). Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proceedings of the National Academy of Sciences USA* **95**, 10396–10401.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *Journal of Molecular Biology* **288**, 147–164.

Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucleic Acids Research* **28**, 228–230.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *Journal of Molecular Biology* **233**, 123–138.

Holm, L. & Sander, C. (1999). Protein folds and families: sequence and structure alignments. *Nucleic Acids Research* **27**, 244–247.

Huang, H., Barker, W. C., Chen, Y. & Wu, C. H. (2003). iProClass: an integrated database of protein family, function and structure information. *Nucleic Acids Research* **31**, 390–392.

Huynen, M. A. & Snel, B. (2000). Gene and context: integrative approaches to genome analysis. *Advances in Protein Chemistry* **54**, 345–379.

Huynen, M., Snel, B., Lathe III, W. & Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Research* **10**, 1204–1210.

Hwang, K. Y., Chung, J. H., Kim, S.-H., Han, Y. S. & Cho, Y. (1999). Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. *Nature Structural Biology* **6**, 691–696.

Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignments and functional genomics. *Proteins: Structure, Function and Genetics* **42**, 378–382.

Iyer, L. M., Aravind, L., Bork, P., Hofmann, K., Mushegian, A. R., Zhulin, I. B. & Koonin, E. V. (2001). Quod erat demonstrandum? The mystery of

experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biology* 2, research0051.1–research0051.11.

Jackson, R. M. & Russell, R. B. (2000). The serine protease inhibitor canonical loop conformation: examples found in extracellular hydrolases, toxins, cytokines and viral proteins. *Journal of Molecular Biology* 296, 325–334.

Jackson, R. M. & Russell, R. B. (2001). Predicting function from structure: examples of the serine protease inhibitor canonical loop conformation found in extracellular proteins. *Computational Chemistry* 26, 31–39.

Jeffery, C. J. (1999). Moonlighting proteins. *Trends in Biochemical Science* 24, 8–11.

Jeffery, C. J., Bahnson, B. J., Chien, W., Ringe, D. & Petsko, G. A. (2000). Crystal structure of rabbit phosphoglucose isomerase, a glycolytic enzyme that moonlights as neuroleukin, autocrine motility factor, and differentiation mediator. *Biochemistry* 8, 955–964.

Jensen, L. J., Gupta, R., Blom, N., Devos, D., Tamames, J., Kesmir, C., Nielsen, H., Staerfeldt, H. H., Rapacki, K., Workman, C., *et al.* (2002). Prediction of human protein function from post-translational modifications and localization features. *Journal of Molecular Biology* 319, 1257–1265.

Jeong, S. S. & Chen, R. (2001). Functional misassignment of genes. *Nature Biotechnology* 19, 95.

Jones, J., Field, J. K. & Risk, J. M. (2002). A comparative guide to gene prediction tools for the bioinformatics amateur. *International Journal of Oncology* 20, 697–705.

Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Research* 30, 42–46.

Karp, R. (1998). What we do not know about sequence analysis and sequence databases. *Bioinformatics* 14, 753–754.

Kasuya, A. & Thornton, J. M. (1999). Three-dimensional structure analysis of PROSITE patterns. *Journal of Molecular Biology* 286, 1673–1691.

Kenny, P. A., Liston, E. M. & Higgins, D. G. (1999). Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene* 232, 11–23.

Kinch, L. N., Wrabl, J. O., Krishna, S. S., Majumdar, I., Sadreyev, R. I., Qi, Y., Pei, J., Cheng, H. & Grishin, N. V. (In Press). CASP5 Assessment of fold recognition target predictions. *Proteins: Structure, Function and Genetics.*

Koehl, P. (2001). Protein structure similarities. *Current Opinion in Structural Biology* 11, 348–353.

Kolesov, G., Mewes, H. W. & Frishman, D. (2001). SNAPping up functionally related genes based on context information: a colinearity-free approach. *Journal of Molecular Biology* 311, 639–656.

Kolesov, G., Mewes, H. W. & Frishman, D. (2002). SNAPper: gene order predicts gene function. *Bioinformatics* 18, 1017–1019.

Lan, N., Jansen, R. & Gerstein, M. (2002). Towards a systematic definition of protein function that scales to the genome level: defining function in terms of interactions. *Proceedings of the IEEE* 90, 1848–1858.

Lan, N., Montelione, G. T. & Gerstein, M. (2003). Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Current Opinion in Structural Biology* 7, 44–54.

Lesk, A. M. (2001). *Introduction to Protein Architecture: The Structural Biology of Proteins.* Oxford: Oxford University Press.

Lesk, A. M. (2002). *Introduction to Bioinformatics.* Oxford: Oxford University Press.

Lesk, A. M. & Chothia, C. (1980). How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *Journal of Molecular Biology* 136, 225–270.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996a). An evolutionary trace method defines binding surfaces common to protein families. *Journal of Molecular Biology* 257, 342–358.

Lichtarge, O., Bourne, H. R. & Cohen, F. E. (1996b). Evolutionarily conserved $G_{\alpha\beta\gamma}$ binding surfaces support a model of the G protein-receptor complex. *Proceedings of the National Academy of Sciences USA* 93, 7507–7511.

Lichtarge, O. & Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology* 12, 21–27.

Lo Conte, L., Brenner, S. E., Hubbard, T. J. P., Chothia, C. & Murzin, A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Research* 30, 264–267.

Madabushi, S., Yao, H., Marsh, M., Kristensen, D. M., Philippi, A., Sowa, M. E. & Lichtarge, O. (2002). Structural clusters of evolutionary trace residues are statistically significant and common in proteins. *Journal of Molecular Biology* 316, 139–154.

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. (1999). A combined algorithm for genome-wide prediction of protein function. *Nature* 402, 83–86.

Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. (1997). MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research* 25, 28–30.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., *et al.* (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research* 31, 315–318.

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Current Opinion in Structural Biology* 8, 380–387.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins

database for the investigation of sequences and structures. *Journal of Molecular Biology* 247, 536–540.

NAGANO, N., ORENGO, C. A. & THORNTON, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology* 321, 741–765.

NANDURKAR, H. H., CALDWELL, K. K., WHISSTOCK, J. C., LAYTON, M. J., GAUDET, E. A. & NORRIS, F. A. (2001). Characterization of an adapter subunit to a phosphatidylinositol (3)P 3-phosphatase: identification of a myotubularin-related protein lacking catalytic activity. *Proceedings of the National Academy of Sciences USA* 98, 9499–9504.

NATALE, D. A., SHANKAVARAM, U. T., GALPERIN, M. Y., WOLF, Y. I., ARAVIND, L. & KOONIN, E. V. (2000). Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biology* 1, research0009.1–0009.19.

NOVICHKOV, P. S., GELFAND, M. S. & MIRONOV, A. A. (2001). Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* 17, 1011–1018.

ORENGO, C. A., BRAY, J. E., BUCHAN, D. W., HARRISON, A., LEE, D., PEARL, F. M., SILLITOE, I., TODD, A. E. & THORNTON, J. M. (2002). The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* 2, 11–21.

OUZOUNIS, C. A., COULSON, R. M. R., ENRIGHT, A. J., KUNIN, V. & PEREIRA-LEAL, J. B. (2003). Classification schemes for protein structure and function. *Nature Reviews Genetics* 4, 508–519.

PARSONS, J. F., LIM, K., TEMPCZYK, A., KRAJEWSKI, W., EISENSTEIN, E. & HERZBERG, O. (2002). From structure to function: YrbI from *Haemophilus influenzae* (HI1679) is a phosphatase. *Proteins: Structure, Function and Genetics* 46, 393–404.

PEARL, F. M., BENNETT, C. F., BRAY, J. E., HARRISON, A. P., MARTIN, N., SHEPHERD, A., SILLITOE, I., THORNTON, J. & ORENGO, C. A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Research* 31, 452–455.

PELLEGRINI, M., MARCOTTE, E. M., THOMPSON, M. J., EISENBERG, D. & YEATES, T. O. E. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences USA* 96, 4285–4288.

PERETZ, D., WILLIAMSON, R. A., LEGNAME, G., MATSUNAGA, Y., VERGARA, J., BURTON, D. R., DEARMOND, S. J., PRUSINER, S. B. & SCOTT, M. R. (2002). A change in the conformation of prions accompanies the emergence of a new prion strain. *Neuron* 34, 921–932.

PERUTZ, M. F. (1983). Species adaptation in a protein molecule. *Molecular Biology and Evolution* 1, 1–28.

PIKE, R. N., BOTTOMLEY, S. P., IRVING, J. A., BIRD, P. I. & WHISSTOCK, J. C. (2002). Serpins: finely balanced conformational traps. *IUBMB Life* 54, 1–7.

PONTING, C. P. (2001). Issues in predicting protein function from sequence. *Briefings in Bioinformatics* 2, 19–29.

PRZYTYCKA, T., AURORA, R. & ROSE, G. D. (1999). A protein taxonomy based on secondary structure. *Nature Structural Biology* 6, 672–682.

RILEY, M. (1993). Functions of gene products of *Escherichia coli*. *Microbiological Reviews* 57, 862–952.

RILEY, M. (1997). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucleic Acids Research* 25, 51–52.

RILEY, M. (1998). Systems for categorizing functions of gene products. *Current Opinion in Structural Biology* 8, 388–392.

RISON, S. C. G., HODGMAN, T. C. & THORNTON, J. M. (2000). Comparison of functional annotation schemes for genomes. *Functional and Integrative Genomics* 1, 56–69.

ROST, B. (2002). Enzyme function less conserved than anticipated. *Journal of Molecular Biology* 318, 595–608.

SCHONBRUN, J., WEDEMEYER, W. J. & BAKER, D. (2002). Protein structure prediction in 2002. *Current Opinion in Structural Biology* 12, 348–354.

SERRES, M. H., GOPAL, S., NAHUM, L. A., LIANG, P., GAASTERLAND, T. & RILEY, M. (2001). A functional update of the *Escherichia coli* K-12 genome. *Genome Biology* 2, research0035.1–0035.7.

SERVANT, F., BRU, C., CARRÈRE, S., COURCELLE, E., GOUZY, J., PEYRUC, D. & KAHN, D. (2002). ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics* 3, 246–251.

SHAH, I. & HUNTER, L. (1997). Predicting enzyme function from sequence: a systematic appraisal. *Proceedings of the International Conference on Intelligent Systems in Molecular Biology* 5, 276–283.

SHAPIRO, L. & HARRIS, T. (2000). Finding function through structural genomics. *Current Opinion in Biotechnology* 11, 31–35.

SIGRIST, C. J., CERUTTI, L., HULO, N., GATTIKER, A., FALQUET, L., PAGNI, M., BAIROCH, A. & BUCHER, P. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics* 3, 265–274.

SKOLNICK, J., FETROW, J. S. & KOLINSKI, A. (2000). Structural genomics and its importance for gene function analysis. *Nature Biotechnology* 18, 283–287.

SMITH, T. F. & ZHANG, X. (1997). The challenges of genome sequence annotation or 'the devil is in the details'. *Nature Biotechnology* 15, 1222–1223.

SMITH, T. F. (1998). Functional genomics – bioinformatics is ready for the challenge. *Trends in Genetics* 14, 291–293.

SOWA, M. E., HE, W., SLEP, K. C., KERCHER, M. A., LICHTARGE, O. & WENSEL, T. G. (2001). Prediction and confirmation of a site critical for effector regulation

of RGS domain activity. *Nature Structural Biology* 8, 234–237.

Sowa, M. E., He, W., Wensel, T. G. & Lichtarge, O. (2000). A regulator of G protein signaling interaction surface linked to effector specificity. *Proceedings of the National Academy of Sciences USA* 97, 1483–1488.

Spiess, C., Beil, A. & Ehrmann, M. (1999). A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein. *Cell* 97, 339–347.

Srinivasaraq, G. Y., Yeh, L. S., Marzec, C. R., Orcutt, B. C. & Barker, W. C. (1999). PIR-ALN: a database of protein sequence alignments. *Bioinformatics* 15, 382–390.

Stein, L. (2001). Genome annotation: from sequence to biology. *Nature Reviews Genetics* 2, 493–503.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. & Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* 29, 22–28.

Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *Journal of Molecular Biology* 208, 1–22.

Teichmann, S. A., Murzin, A. G. & Chothia, C. (2001a). Determination of protein function, evolution and interactions by structural genomics. *Current Opinion in Structural Biology* 11, 354–363.

Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001b). The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *Journal of Molecular Biology* 311, 693–708.

Teichmann, S. A., Rison, S. C., Thornton, J. M., Riley, M., Gough, J. & Chothia, C. (2001c). Small-molecule metabolism: an enzyme mosaic. *Trends in Biotechnology* 19, 482–486.

Thornton, J. M. (2001). From genome to function. *Science* 292, 2095–2097.

Thornton, J. M., Orengo, C. A. & Pearl, F. M. (1999). Protein folds, functions and evolution. *Journal of Molecular Biology* 293, 333–342.

Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of protein function, from a structural perspective. *Journal of Molecular Biology* 307, 1113–1143.

Todd, A. E., Orengo, C. A. & Thornton, J. M. (2002). Plasticity of enzyme active sites. *Trends in Biochemical Sciences* 27, 419–426.

Tramontano, A. (2003). Of men and machines. *Nature Structural Biology* 10, 87–90.

Vlahovicek, K., Murvai, J., Barta, E. & Pongor, S. (2002). The SBASE protein domain library, release 9.0: an online resource for protein domain identification. *Nucleic Acids Research* 30, 273–275.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P. & Snel, B. (2003). STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research* 31, 258–261.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Science* 5, 1001–1013.

Whisstock, J., Skinner, R. & Lesk, A. M. (1998). An atlas of serpin conformations. *Trends in Biochemical Sciences* 23, 63–67.

Wilks, H. M., Hart, K. W., Feeney, R., Dunn, C. R., Muirhead, H., Chia, W. N., Barstow, D. A., Atkinson, T., Clarke, A. R. & Holbrook, J. J. (1988). A specific, highly active malate dehydrogenase by redesign of a lactate dehydrogenase framework. *Science* 242, 1541–1544.

Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *Journal of Molecular Biology* 297, 233–249.

Wistow, G. & Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Science* 236, 1554–1556.

Wu, G., Fiser, A., ter Kuile, B., Sali, A. & Müller, M. (1999). Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proceedings of the National Academy of Sciences USA* 96, 6285–6290.

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S. M. & Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* 30, 303–305.

Yao, H., Kristensen, D. M., Mihalek, I., Sowa, M. E., Shaw, C., Kimmel, M., Kavraki, L. & Lichtarge, O. (2003). An accurate, sensitive, and scalable method to identify functional sites in protein structures. *Journal of Molecular Biology* 326, 255–261.

Yona, G., Linial, N. & Linial, M. (2000). ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Research* 28, 49–55.

Zhang, H., Huang, K., Li, Z., Banerjei, L., Fisher, K. E., Grishin, N. V., Eisenstein, E. & Herzberg, O. (2000). Crystal structure of YbaK protein from *Haemophilus influenzae* (HI1434) at 1·8 Å resolution: functional implications. *Proteins: Structure, Function and Genetics* 40, 86–97.

Zhang, C. & Kim, S. H. (2003). Overview of structural genomics: from structure to function. *Current Opinion in Chemical Biology* 7, 28–32.